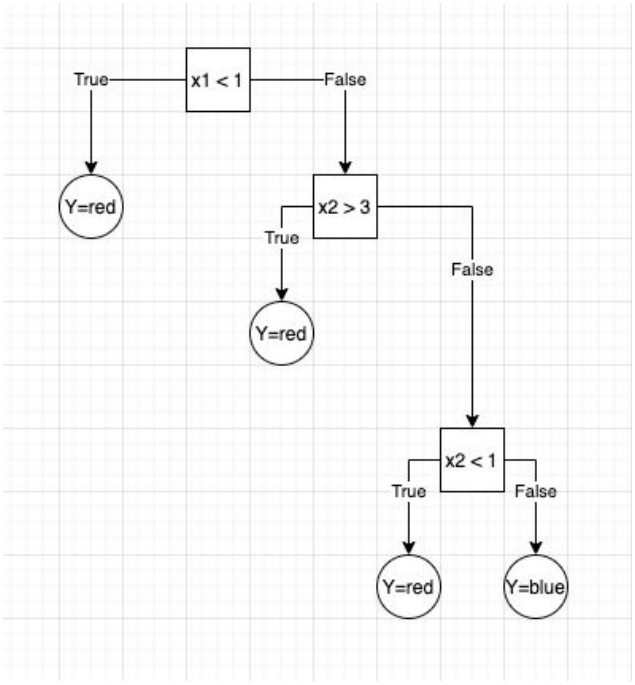
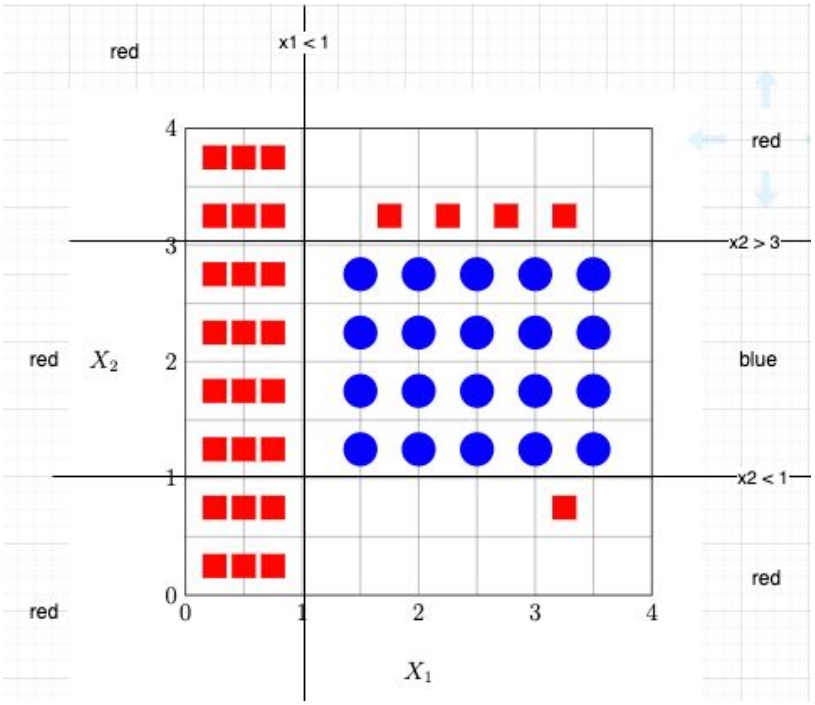
1.

a. A decision tree is sequentially partitioning the dataset using one feature at a time. To construct a single (or multiple for complex splits) feature is picked and then partition the data based on a threshold τ. A feature can be used multiple times to partition the dataset. The purity of a split can be measured using the gini index, from this you pick the feature and threshold to minimize.

b.



c.



2.

a.

i. this ignores the penalty all together as the 2nd summation notation would always be multiplied by 0 and therefore equal 0.

ii. this scenario puts an incredible amount of weight on the penalty portion and there would be few scenarios where this would output non-trivial results.

b.

i. if p=0 the penalty term would be multiplying lambda by 5 and nothing else.

ii. 2

iii.

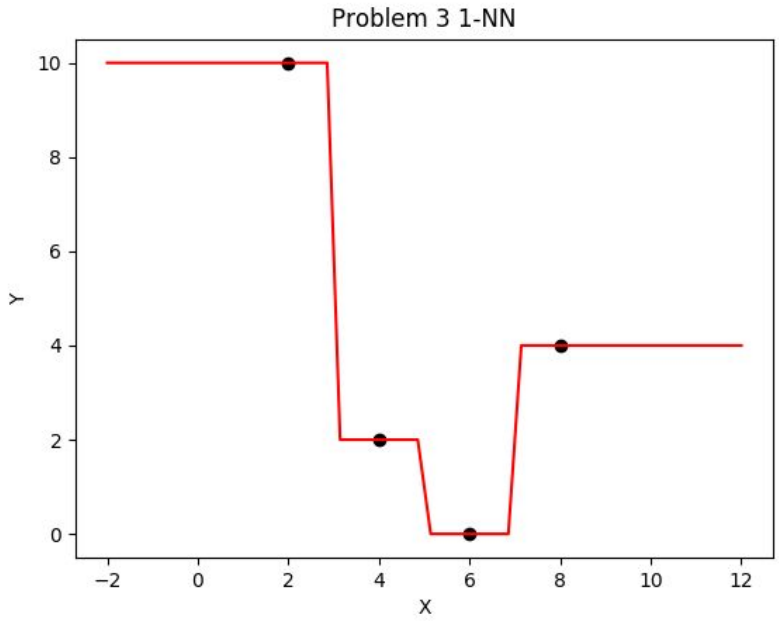
1. greedy search: start with a set of all available features and try removing one at a time and comparing the (total loss + model complexity) to previous tries.
2. fit all subclasses with a single feature: make a model using the features that performed the best individually (up to a threshold). This is the fastest but can fall prey to features with high correlation.

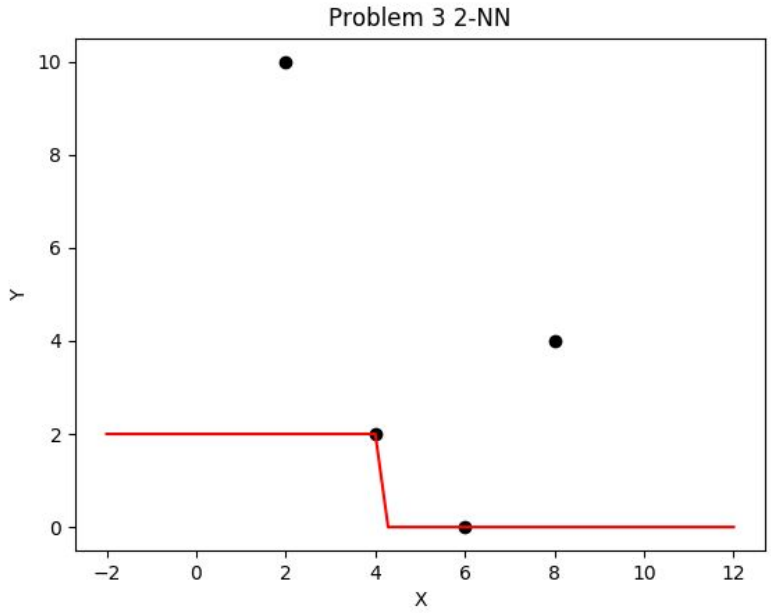
c. 3

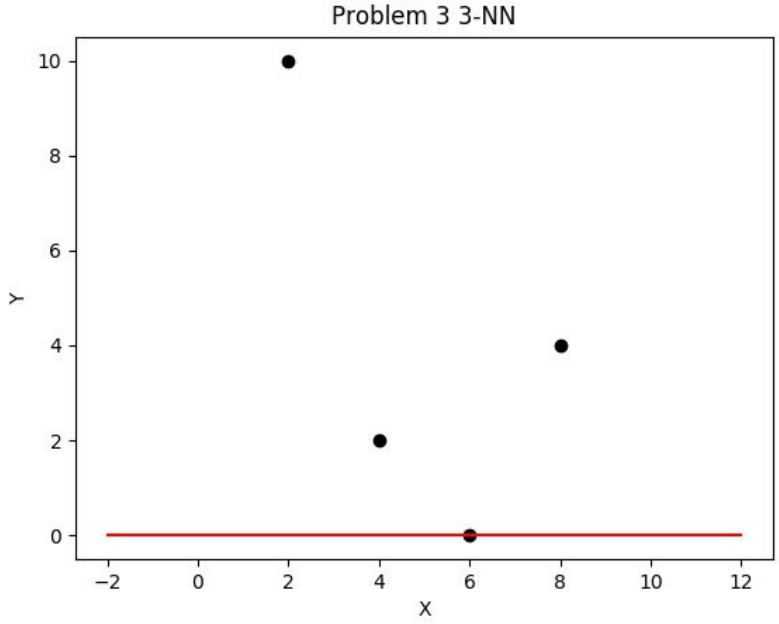
d. p=1 will favor smaller coefficients as the Y(i) term will scale quicker than the penalty term. p=2 will favor greater coefficients because the value of the penalty term will outpace the first summation term. As lambda increases both of these preferences are amplified.

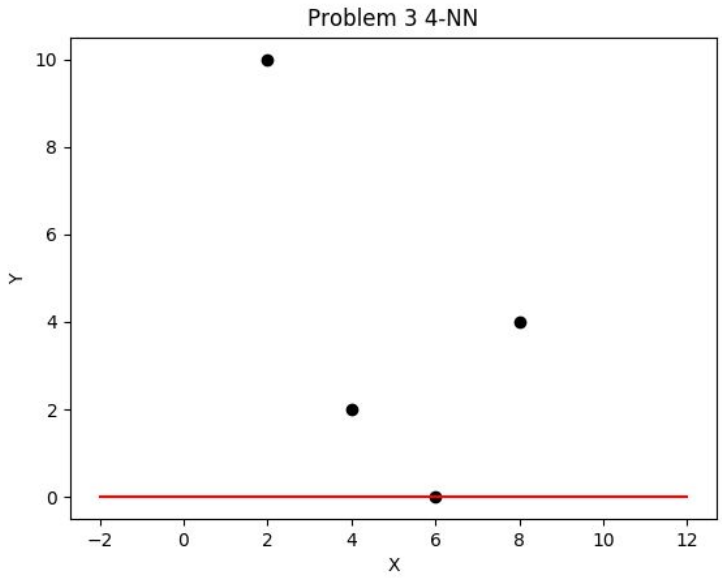
3. program that generated these graphs in zip folder titled "q3.py"

- 1nn: absolute difference of points
- 2..4nn: as knn increases the curve smoothes out. The graph goes to the value of the k nearest neighbors.




Problem 3 2-NN
10
8
6
4
2
0
Y
−2
0
2
4
6
8
10
12
X


Problem 3 3-NN
10
8
6
4
2
0
Y
−2
0
2
4
6
8
10
12
X



4.

- Overfitting: the model predicts a dataset too well. This negatively effects the models ability to generalize.
- Underfitting: model cannot generalize about training or new datasets. The polynomial degree is to "rigid" and is not recognizing correlations that could be valuable for future data extrapolation.
- Undesirable: they both will fail to make non-trivial or rather inaccurate predictions about future data which is the main purpose of these models.
- high bias + low variance = underfitting, i.e. if the model is not performing well on the training set it is likely ignoring key data points and won't be able to accurately predict outcomes on new data. low bias + high variance = overfitting, i.e. if the model is performing to well on the training data (memorizing the data) but fails to make valuable predictions on new data.

5. Used for dimension reduction and to filter noise out of datasets, an example is image compression. PCA is unsupervised but linear regression is supervised. Because of this PCA uses no dependant variables and linear regression does.

6.

a. It would form verticle line at $x_1 = 0$. LDA is better suited for smaller data sets therefor with a total dataset of 2000 the single outlying red x would not notably impact the decision boundary.
b. A mostly vertical line would form at $x_1 = 0$. The QDA decision boundary would be impacted by the outlying red x so there would be a small bump toward negative x_1.
c. SVM is impacted the most of the 3 by outliers. The boundary would be at $x_1 = -2$ and the "black" margin would be from -5 to -2 and the "red" margin would be from -2 to 1. The decision boundary saw such a large shift from the previous 2 classifiers because SVM splits the data and then maximizes the margins on either side.
d. The boundary would gradually shift more toward $x_1 = 0$ and the "black" margin would grow as the "red" shrank as less and less weight was put on the single outlying red x.